SSIM-Based End-to-End Distortion Modeling for H.264 Video Coding*

Yuxia Wang, Yuan Zhang, Rui Lu, and Pamela C. Cosman*

[1] Communication University of China, Beijing, 100024, China
[2] University of California, San Diego, CA, 92093-0407, USA

Abstract. The estimation of end-to-end distortion plays a key role in error-resilient video coding and perceptual quality control. The traditional end-to-end distortion estimation methods are mainly based on the MSE or MAD values, which sometimes poorly reflect subjective perception. This paper proposes a novel method to model the end-to-end quality degradation based on the SSIM index. Using factors extracted from the encoder, we build the models by considering the source distortion, the error-propagated distortion and the error-concealment distortion. These models can be used in joint source-channel coding with rate-distortion optimization as well as error-resilient video coding based on perception.

Keywords: end-to-end distortion, error propagation, quality evaluation, GLM, SSIM.

1 Introduction

When compressed videos are transmitted through error-prone networks, the video quality at the receiver side can be highly affected by packet losses in addition to compression artifacts. In a packet-loss environment, transmitting the hybrid-coded video such as H.264 often suffers from error propagation, which may lead to the well-known drifting phenomenon [1]. Fortunately, estimating the end-to-end distortion in the error-resilient coding of the encoder can help to control the errors.

A generalized end-to-end approach has been proposed for video communication over packet-switched networks [2], in which a set of global distortion metrics were derived in terms of MAD. Further, a recursive optimal per-pixel estimate (ROPE) algorithm was proposed to estimate the end-to-end distortion at the pixel level [3]. Zhang et al. [4] proposed a concise and efficient end-to-end distortion model, in which the overall distortion is categorized into source, error-propagated and error-concealment distortion items.

All the previous models calculate the distortions based on MSE or MAD. However, sometimes these objective measurements poorly reflect perceptual quality, especially for error-concealed videos. Recently, many objective metrics were

* This reserch has been supported in part by National Natural Science Foundation of China: 61001177.

W. Lin et al. (Eds.): PCM 2012, LNCS 7674, pp. 117–128, 2012.

proved effective for video quality evaluation, such as SSIM [5], JND [6] and VQM [7] etc. Due to the accuracy of these full-reference metrics, many new methods of quality evaluation or prediction are proposed based on them in terms of different applications. In particular, a theoretical framework for end-to-end video quality prediction for MPEG sequences based on the SSIM index was proposed in [8]. The temporal variations of video quality also have different effects on global video quality, which is explored using the PSNR and the SSIM indices in [9]. A network-based model for video packet importance based on the VQM index is proposed in [10], which considers the effects of both compression artifacts and packet losses. Some rate-distortion (R-D) optimized techniques for video coding have been proposed based on the SSIM metric recently [11, 12], yet the channel distortion is not considered.

SSIM is widely considered to be more reasonable in measuring the perceptual visual quality than the MSE. We aim to apply the SSIM measurement in the error-resilient video coding. The challenge is how to estimate the SSIM index at the encoder side without the decoded lossy video. In this paper, we propose a model to predict the end-to-end perceptual quality degradation for compressed videos transmitted in an error-prone environment. We explore the relationship between the errors in the pixel domain and the visual quality degradation at the decoder side, and then predict the perceptual quality scores (SSIM) based on factors extracted from the encoder. We consider the quality scores at the macroblock (MB) level so that the perceptual quality of each MB can be improved accurately.

The organization of this paper is as follows. Section 2 describes the idea of the end-to-end distortion estimation at the MB level, based on the algorithm of the SSIM index. In Section 3, the test sequences and coding configuration are given, and a new method to model the end-to-end quality degradation of H.264 videos at the MB level is proposed. Section 4 presents the final models, followed by the analysis and validation of the experiments. Section 5 concludes the paper.

2 End-to-End Quality Evaluation at the MB Level

Based on the previous method [4], we take the end-to-end distortion as the comprehensive effect of the source distortion, the error propagation distortion and the error-concealment distortion, in which the estimate of pixel distortion is derived by simulating the decoding process multiple times in the encoder. The problem is that MSE or MAD value is not always consistent with the perceptual quality degradation.

Wang et al. proposed the SSIM index [5] to measure the subjective similarity between the original and distorted images. The SSIM index for a still image is derived based on similarities of local luminance, contrast and structure between a reference image and a distorted image :

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{1}$$

where μ_x and μ_y are the means of the luminance values of the original block x and the reconstructed block y, and σ_x and σ_y are the standard deviations. σ_{xy} is the cross correlation, and C_1 and C_2 are constants used to avoid instability when the means and variances are close to zero. In implementation, a sliding window moves pixel by pixel horizontally and vertically through all the rows and columns of the image until the bottom-right corner is reached. The SSIM index of the whole image is obtained by averaging the local SSIM indices calculated using a sliding window of 11×11 .

To calculate the SSIM quality score for each MB, we have conducted the experiment with different window sizes and different block sizes. The statistics show that a 11×11 sliding window within a MB balances the subjective perception and computational complexity very well. The values of SSIM are in the range [0, 1], where 0 corresponds to the worst quality, and 1 is the best quality. Considering the influence of the chroma, the final SSIM index is obtained by formula (2):

$$SSIM = 0.8SSIM_Y + 0.1SSIM_{CR} + 0.1SSIM_{CB} \tag{2}$$

Where $SSIM_Y$ denotes the SSIM index values of the luminance component, $SSIM_{CR}$ and $SSIM_{CB}$ denote the SSIM values of the color components.

3 End-to-End Distortion Modeling

3.1 Test Sequences and Coding

The SSIM-based end-to-end distortion is modeled based on six coded video sequences (Foreman, News, Akyio, Coastguard, Hall and Mobile) which are selected according to various levels of detail and motion types. To cover various source distortions, we encode each sequence with six different QP values (20, 24, 28, 32, 36, 40). H.264 JM10.2 is adopted as the encoder, with the coding conditions shown in Table 1.

Table 1. Coding and sequence parameters

	settings
Spatial resolution	QCIF(176*144)
Duration (frames)	100
Compression standard	H.264
GOP structure	IPPP...
Frame rate	30
QP values	20, 24, 28, 32, 36, 40
Rate control	off
Packet losses	0.03, 0.05, 0.1, 0.2
Concealment	frame copy

To explore the effect of packet losses on the quality of decoded videos, we randomly drop packets from each sequence until the target packet loss rate is obtained. We assume the packets of the first frame are conveyed reliably. For each packet loss rate, each H.264 stream will be decoded as 30 sequences for 30 realizations of the lossy channel. The mean of the quality scores of the 30 sequences is calculated to build the model.

3.2 Modeling Approach

In order to predict the MB-level SSIM scores of the videos, we use a generalized linear model (GLM) [13], which can be represented as:

$$g(p) = \gamma + \sum_{j=1}^{P} x_j \beta_j \tag{3}$$

where $g(.)$ is the link function which is chosen depending on the distribution. The parameter p is modeled as a function of P factors (x_j), which denotes the SSIM score we are trying to predict. γ is the constant term, and $\beta_1, \beta_2, \ldots \beta_P$ are the coefficients of the factors. The coefficients and the constant term are usually unknown and need to be estimated from the data. Given N observations, one can fit models using up to N parameters. The simplest model (Null model) has only one parameter: the constant γ. On the other hand, a full model can have as many factors as observations.

Five factors are extracted from each MB of each frame at the encoder side:

(1) D_s: the estimated source distortion of each encoded MB.

(2) D_{ep}: the estimated error-propagated distortion of each MB from the reference frame.

(3) D_{ec}: the error-concealment distortion for each MB.

(4) *Qstep*: the quantization step (six values) for each sequence.

(5) *PLR*: the setting of packet loss rate (four values) for each sequence.

D_s, D_{ep} and D_{ec} are calculated based on the method in [4], and the last two items can be recursively calculated after a frame has been encoded. As discussed in the literature, the estimation algorithm of D_{ec} depends on the method of error concealment at the decoder side. To reduce the computation complexity, here we use the method of frame copy as referred in the JM decoder. The overall end-to-end distortion of a block can be taken as the sum of that from each pixel. We extend the block size from 4×4 to 16×16 by summation, as we aim at predicting the video quality at the MB level.

Figure 1 gives the distributions of the first three factors and the SSIM scores for the sequence Foreman when *PLR* equals 20%. Similar distributions are obtained for other conditions of all the sequences. We use "log" as the link function in model building based on the approximately Poisson distribution of SSIM scores in Figure 1, where the expression of the equation is

$$g(p) = \log(p) \tag{4}$$

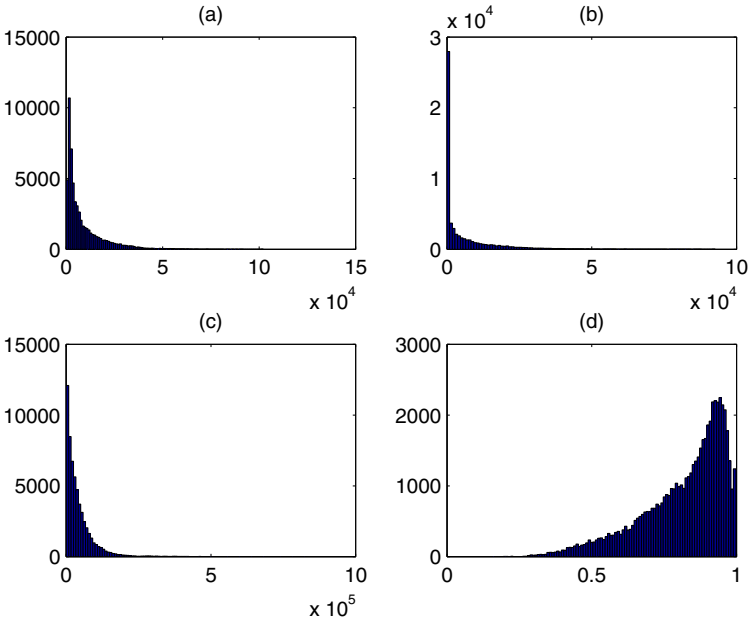

Fig. 1. The histogram distributions of (a) D_s (b) D_{ep} (c) D_{ec} (d) SSIM scores

Four-fold cross-validation is used to determine a model of the right size. In cross-validation, the data set is divided into four parts. Then three parts are used to train the model, and the resulting model is tested on the fourth part which was left out of the training. This process is repeated four times, with a different part left out each time. Factors are added into a model in order of importance. The MATLAB function "sequentialfs" is used, which performs sequential feature selection. It selects factors by sorting the importance from all the factors, based on the mean squared error between predicted values and actual values. The selection proceeds until there is no improvement in prediction.

4 Experiment Results

4.1 Final Models and the Analysis

As the first exploration, we would like to see if the end-to-end estimation function in [4] is still valid for SSIM scores instead of MSE. So D_s, D_{ep}, D_{ec} and PLR are selected as the factors, and the interaction terms are also added in model-building. With six sequences of 100 frames (99 P frames), we set six QP values and four $PLRs$ as shown in Table 1. There are 99 MBs per frame. We extract the factors for each MB, so there are more than 1 million data totally ($N = 6 \times 6 \times 4 \times 99 \times 99 = 1411344$). We randomly select one-tenth of the data to build the models in case of overtraining.

Figure 2 shows on the y-axis the correlation between the SSIM scores predicted by our model and the actual SSIM scores. Here each actual SSIM score is the mean of thirty SSIM scores from the thirty lossy channels given a certain PLR. The number of factors included in the model is on the x-axis, in order of importance. Table 2 gives the factors in order of importance. We can see that D_s is the most important factor which makes the correlation increase significantly. It is reasonable that the source distortion due to quantization is the main cause for quality degradation of the decoded video. PLR*D_s and PLR*D_{ec} are also useful factors in the model with negative coefficients, which means larger values lead to lower SSIM scores and worse quality of the MBs. The impact of packet losses will grow bigger with the increasing PLR as expected. Unlike the conclusion in the literature [4], D_{ep} is not an important factor and is excluded in the final model. The reason is that for an intra MB or a skip MB, there is no distortion of error propagation from quantization during encoding, D_{ep} will be zero, as a refreshed value. While the distortion from packet loss results in the main quality degradation, and the recursive estimating calculation of D_{ec} has reflected the quality impairment of the error propagation.

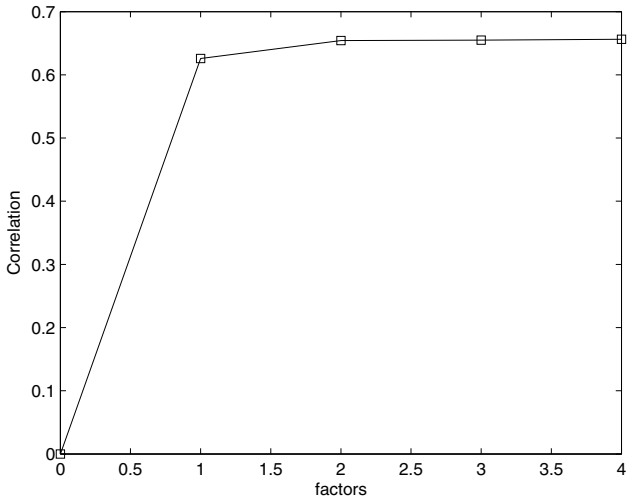

Fig. 2. Correlation between predicted scores and actual SSIM

The first model is built for all MBs, including intra, inter and skip ones. However, the characteristics of the three types of MBs are quite different. As concluded in paper [14], the $SSIM - Qstep$ relation model of I frames in H.264 can be modeled accurately with a linear function. That means the SSIM score of an intra MB can be predicted simply using one factor: $Qstep$. Therefore, we carry out this experiment only for Inter MBs. Since $Qstep$ can reflect the source

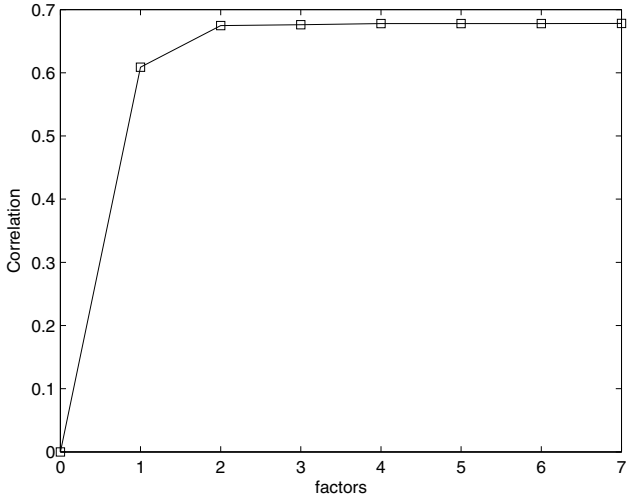

Fig. 3. Correlation between predicted scores of the new model and actual SSIM

distortion as a known parameter in the encoder, we replace D_s with $Qstep$ to build the new model. Figure 3 gives, on the y-axis, the correlation between the predicted SSIM scores and the actual SSIM scores. Table 3 gives the factors and corresponding coefficients in order of importance. We can see that $Qstep$ plays a key role in this model presenting the same impact as D_s. PLR*D_{ec} is the second important factor having a similar effect as in the first model. D_{ep} is also included with the negative coefficient, which means there is the distortion of error propagation from quantization error for an inter MB besides the error-propagated distortion from packet losses. With these factors ($Qstep$, D_{ec}, D_{ep} and PLR), this model gives a better predictive result with lower computational complexity. Also we can simplify this model further by excluding the last five factors since the improvement on the correlation is not evident.

Table 2. Factors in order of importance for Model-ALL

Order	Factors	Coefficients
Intercept	1	-7.54e-002
1	D_s	-4.27e-006
2	PLR*D_s	-5.19e-006
3	PLR*D_{ec}	-1.78e-006
4	D_{ec}	3.13e-007

Table 3. Factors in order of importance for Model-INTER

Order	Factors	Coefficients
Intercept	1	4.28e-001
1	$Qstep$	-5.10e-001
2	PLR*D_{ec}	-8.85e-006
3	D_{ep}	-1.78e-006
4	PLR*$Qstep$	-4.22e-001
5	PLR	2.98e-001
6	D_{ec}	1.24e-007
7	PLR*D_{ep}	3.38e-006

4.2 Validation of the Proposed Models

To validate the proposed models, we perform the experiment with another two sequences: Suzie and Container, which are not included in the model-building process. The videos are encoded by randomly setting a coding parameter QP with JM10.2 encoder. Similarly, given a PLR, a compressed stream suffering from packet losses by 30 channels is decoded as 30 lossy videos. Figure 4 shows the decoded videos from one channel. The actual SSIM score of each MB from the video is calculated by averaging the 30 lossy versions from 30 channel realizations. On the other hand, using the proposed model (Model-ALL), we can predict the SSIM score of the same MB by means of the effective factors as shown in Table 2.

Figures 5 and 6 show the estimated and the actual SSIM scores for video Suzie when QP is 40 and PLR is 5%. Figures 7 and 8 give the comparison for video Container when QP is 36 and PLR is 20%. To demonstrate the varying SSIM scores clearly, we randomly pick two segments of 200 MBs (about 2-frame-length for QCIF format). We can see that there are good correlations between the estimated SSIM scores by our models and the actual SSIM scores for both videos. Note that the distributions of the SSIM scores change with the different

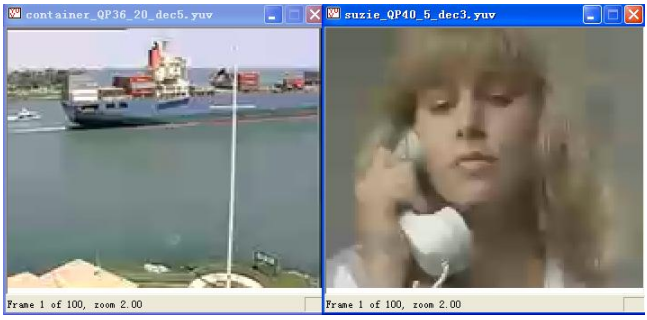

Fig. 4. The decoded lossy videos: Suzie and Container

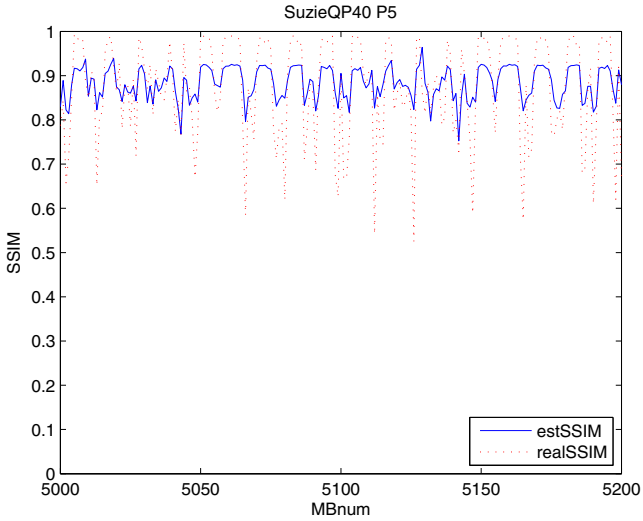

Fig. 5. The estimated SSIM scores vs. the actual SSIM scores of 200 MBs for video Suzie

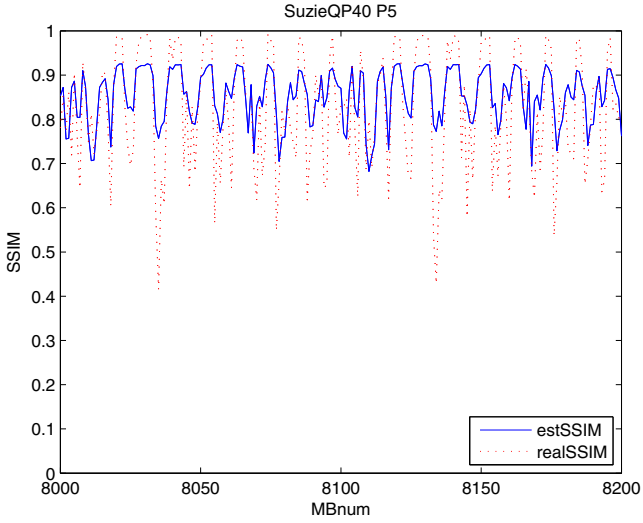

Fig. 6. The estimated SSIM scores vs. the actual SSIM scores of another 200 MBs for video Suzie

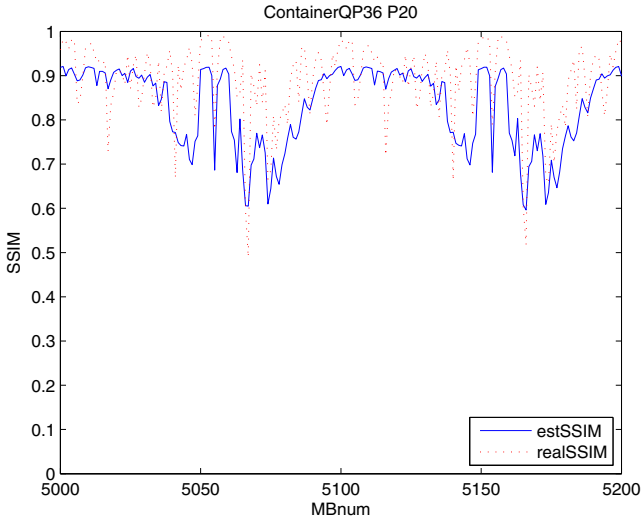

Fig. 7. The estimated SSIM scores vs. the actual SSIM scores of 200 MBs for video Container

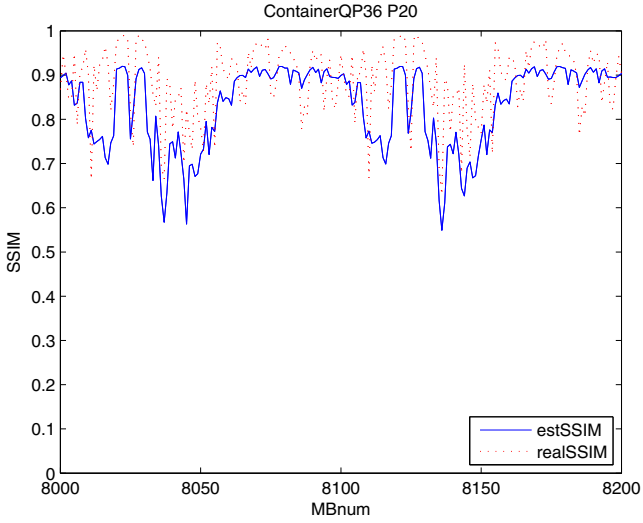

Fig. 8. The estimated SSIM scores vs. the actual SSIM scores of another 200 MBs for video Container

video contents as expected. For Suzie, the motion is well-distributed in the whole frame, so the quality scores fluctuate in a certain range, quite regularly. The quality scores change with an approximate period of 11 MBs, that means the quality change of each horizontal row of MBs is quite similar as expected. While there are some evident drops for the quality scores of Container as shown in Figures 7 and 8. There is a periodicity about 100 MBs for these scores besides the slight local fluctuations. The reason is that, given a certain QP and one PLR, the quality degradation of videos happens mainly in the regions with more details or large motion, i.e. the moving ship in this picture. Thus the MBs in this region get lower scores while the MBs of other regions (like water surface) get higher scores. Similar results can be achieved for other settings of QP and PLR. In short, the proposed model can predict the end-to-end distortion of the videos well. Although the estimated SSIM scores can not reach the peak or valley points of the actual scores, as shown in Figures 5 and 6, the variation trend is very similar. That means the dynamic range of SSIM values is compressed to some extent. However, our model is supposed to be used in joint source-channel rate-distortion optimization and error-resilient video coding. On this condition, we consider the relative quality of a MB given different coding modes, rather than the absolute scores. So the model is valid to predict and control the video quality at the MB level.

5 Conclusion

A novel method of modeling the end-to-end video quality degradation at the MB-level has been proposed in this paper. Considering the varying QP values and multiple packet loss rates, we have built GLM models to predict the perceptual quality degradation of H.264 videos based on the SSIM index. The source distortion, the error-propagated distortion and the error concealment distortion contribute to the actual quality degradation at the decoder side, especially for inter MBs. By estimating these distortions in the encoder, given a certain packet loss rate, we can predict the end-to-end quality using our models. The experiment results show that the proposed models give a good performance on perceptual video quality estimation. These models can be used to improve the perceptual quality of videos in joint source-channel rate-distortion optimization and error-resilient coding.

References

1. Stuhlmuller, K., Farber, N., Link, M., Girod, B.: Analysis of video transmission over lossy channels. IEEE J. Select. Areas Commun. 18, 1012–1032 (2000)
2. Wu, D., Hou, Y.T., Li, B., Zhu, W., Zhang, Y.-Q., Chao, H.J.: An end-to-end approach for optimal mode selection in Internet video communication: theory and application. IEEE J. Select. Areas Commun. 18(6), 977–995 (2000)
3. Zhang, R., Regunathan, S.L., Rose, K.: Video coding with optimal inter/intra-mode switching for packet loss resilience. IEEE J. Select. Areas Commun. 18(6), 966–976 (2000)

4. Zhang, Y., Gao, W., Lu, Y., Huang, Q., Zhao, D.: Joint source-channel rate-distortion optimization for H.264 video coding over error-prone networks. IEEE Trans. on Multimedia 9(3), 445–454 (2007)
5. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image Quality Assessment: From Error Visibility to Structural Similarity. IEEE Trans. Image Processing 13(4), 600–612 (2004)
6. ATIS: Objective Perceptual Video Quality Measurement Using a JND Based Full Reference Technique. Alliance for Telecommunications Industry Solutions Technical Report, T1.TR. 75-2001 (2001)
7. Pinson, M., Wolf, S.: A new standardized method for objectively measuring video quality. IEEE Trans. on Broadcasting 50(3), 312–322 (2004)
8. Koumaras, H., Kourtis, A., Lin, C.-H., Shieh, C.-K.: A Theoretical Framework for End-to-End Video Quality Prediction of MPEG-based Sequences. In: Proc. 3rd Int'l Conf. on Networking and Services, Athens, Greece (2007)
9. Yim, C., Bovik, A.C.: Evaluation of temporal variation of video quality in packet loss networks. Signal Processing: Image Communication, 24–38 (2011)
10. Wang, Y., Lin, T.-L., Cosman, P.: Network-based model for video packet importance considering both compression artifacts and packet losses. In: IEEE Globecom 2010 (2010)
11. Ou, T.-S., Huang, Y.-H., Chen, H.H.: SSIM-Based Perceptual Rate Control for Video Coding. IEEE Trans. on Circuits and Systems for Video Technology 21(5) (2011)
12. Wang, S., Rehman, A., Wang, Z., Ma, S., Gao, W.: Rate-SSIM Optimization For Video Coding. In: IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP (2011)
13. McCullagh, P., Nelder, J.A.: Generalized Linear Models, 2nd edn. Chapman and Hall (1989)
14. Cui, Z., Zhu, X.: Subjective Quality Optimized Intra Mode Selection for H.264 I Frame Coding Based on SSIM. In: The Sixth International Conference on Image and Graphics (2011)